Robert C. Maher

Audio Forensic Examination

Authenticity, enhancement, and interpretation

The field of audio forensics involves many topics familiar to the general audio digital signal processing (DSP) community, such as speech recognition, talker identification, and signal quality enhancement. There is potentially much to be gained by applying modern DSP theory to problems of interest to the forensics community, and this article is written to give the DSP audience some insight into the types of problems and challenges that face practitioners in audio forensic laboratories. However, this article must also present several of the frustrations and pitfalls encountered by signal processing experts when dealing with typical forensic material due to the standards and practices of the legal system.

Digital Object Identifier 10.1109/MSP.2008.931080

INTRODUCTION

Audio forensics refers to the acquisition, analysis, and evaluation of audio recordings that may ultimately be presented as admissible evidence in a court of law or some other official venue. Audio forensic evidence is typically obtained as part of a civil or criminal law enforcement investigation or as part of the official inquiry into an accident or other civil incident.

The principal concerns of audio forensics are i) establishing the authenticity of audio evidence [1], [2], ii) performing enhancement of audio recordings to improve speech intelligibility and the audibility of low-level sounds [3], [4], and iii) interpreting and documenting sonic evidence, such as identifying talkers, transcribing dialog, and reconstructing crime or accident scenes and timelines [5]. The requirements of admissibility into a court of law mean that the techniques employed must

1053-5888/09/$25.00©2009IEEE

previously have been proven to the court to be unbiased, to have known reliability statistics, to be nondestructive, and to be widely accepted by experts in the field.

For example, a modern digital speech enhancement technique may seem like an obvious choice for cleaning up a recorded surveillance recording prior to preparing a transcript, but the court may need to be convinced that the "enhancement" could not have resulted in a change to the meaning or interpretation of the recorded dialog. If noise is removed by the speech enhancement procedure, the defense and prosecuting attorneys and the court may be concerned that an underlying phoneme was also inadvertently (or deliberately!) altered. For instance, the court could reasonably worry that a signal processing technique might change the interpretation of a noisy utterance "I didn't do it!" to the phrase "I did too do it!", with the obviously different implications for the transcript. Similarly, establishing the proper chain of custody and authenticity when dealing with digital audio files typically goes well beyond the standard operating procedure of academic DSP research labs. Thus, due to these special legal considerations, this article must include some of the arcane history and practices of audio forensics even though the topics and techniques seem out-of-date to those of us in the signal processing research field.

FORENSIC AUDIO RECORDINGS TYPICALLY SUFFER FROM NOISE, DISTORTION, INTERFERING SOUNDS, AND OTHER SIGNAL PROCESSING CHALLENGES THAT CAN IMPEDE PROPER ANALYSIS.

HISTORY

Over the preceding 40 years, forensic audio examination has gradually become a recognized profession. Prior to the availability of DSP tools, most forensic audio analysts worked exclusively with analog magnetic tape and basic recording studio electronics such as analog filters, variable-speed playback equipment, gain compressors, and test equipment such as oscilloscopes, microscopes, and voice spectrographs [3].

Since the 1960s, the U.S. Federal Bureau of Investigation (FBI) has conducted examination of audio recordings for speech intelligibility enhancement and authentication [3]. Law enforcement organizations around the world also have developed procedures and standards for handling audio forensic material in the context of local legal rules and customs.

The authenticity of analog tape recordings was (and still is) assessed using magnetic development, which uses a colloidal suspension of fine magnetic particles in a fluid that evaporates after being spread on the tape [1]. When dry, the magnetic particles stick to the magnetic domains recorded on the tape, thereby revealing visibly under a microscope the underlying magnetic patterns. Determining authenticity from modern digital files can be problematic, as described later in this article.

In predigital days, the forensic audio examiner tasked with the problem of signal enhancement for speech transcription and speaker identification would play the original tape (or a reference copy) repeatedly while adjusting the filter and gain settings to get the best subjective result.

AUDIO FORENSICS AND THE LAW

In the United States, the controlling legal case establishing the admissibility of audio forensic evidence in the form of recorded conversations is generally regarded to be the 1958 ruling in United States versus McKeever (169 F.Supp. 426, 430, S.D.N.Y. 1958). The judge in the McKeever case had to rule on the admissibility of a tape-recorded conversation involving the defendant. Although in the McKeever case a written transcript was ultimately presented to the jury rather than playback of the tape itself, the judge's ruling established a set of requirements that are now used, with some variation, in most state and federal courts in the United States when considering the authenticity of audio recordings. The seven tenets of audio authenticity from the McKeever case are listed in Table 1.

Tenets 1 and 2 are now considered less significant than in the late 1950s when tape recorders were not the ubiquitous and familiar devices judges and juries know of today. Tenets 3 and 4 tend to imply the involvement of audio experts who can examine the physical tape and the subtle magnetic characteristics of the recording device to determine if any splices, stop/start sequences, overrecordings, or other issues exist that would indicate inadvertent damage or deliberate tampering. Tenet 5 demands a proper chain of custody of the recording, while tenet 6 requires that the participants in the recording be identified either by voice or by corroborating witnesses to the recording. Finally, tenet 7 requires that the recorded conversation be spontaneous and not coerced.

In a legal sense, both the authenticity of the physical evidence, consisting of the tape and recording system, and the legal implications of the transcript—which is often subject to interpretation and dispute—are issues for the court to sort out. For example, the fact that a recorded conversation is generally obtained out of court, the participants are not sworn, and witnesses may or may not be available for cross-examination, makes it necessary for the court to determine whether or not the recording is admissible as evidence in a trial.

[TABLE 1] MCKEEVER CASE REQUIREMENTS FOR AUDIO AUTHENTICITY.

1) THAT THE RECORDING DEVICE WAS CAPABLE OF TAKING THE CONVERSATION NOW OFFERED IN EVIDENCE.
2) THAT THE OPERATOR OF THE DEVICE WAS COMPETENT TO OPERATE THE DEVICE.
3) THAT THE RECORDING IS AUTHENTIC AND CORRECT.
4) THAT CHANGES, ADDITIONS, OR DELETIONS HAVE NOT BEEN MADE IN THE RECORDING.
5) THAT THE RECORDING HAS BEEN PRESERVED IN A MANNER THAT IS SHOWN TO THE COURT.
6) THAT THE SPEAKERS ARE IDENTIFIED.
7) THAT THE CONVERSATION ELICITED WAS MADE VOLUNTARILY AND IN GOOD FAITH, WITHOUT ANY KIND OF INDUCEMENT.

EXPERT WITNESSES

In the United States, the various state and federal jurisdictions apply a variety of standards for admitting the testimony of topical experts. The standards are commonly based on the 1923 Frye case (Frye versus United States, 54 App. D.C. 46, 293F.1013, DC Ct App 1923), the Daubert case (Daubert versus Merrell Dow Pharmaceuticals, 509 U.S. 579 1993) or some similar interpretation. The Frye standard requires that the methods and techniques of the expert be generally accepted by the scientific community. Daubert uses the Federal Rules of Evidence to support an acceptability test of relevance and scientific reliability of the expert's testimony. Subsequent cases, such as Kumho Tire Co. versus Carmichael (526 U.S. 137 1999), have extended the Daubert standards beyond scientific testimony to technical and other specialized knowledge (such as audio engineering).

THE 18½-MINUTE GAP

The watershed event for audio forensics was arguably the 1974 investigation of a White House conversation between U.S. President Richard M. Nixon and Chief of Staff H.R. Haldeman recorded in the Executive Office Building in 1972. Investigators discovered that the audio recording contained an unexplained section lasting 18½ minutes during which buzz sounds could be heard but no discernable speech sounds were present. Due to the highly specialized nature of the technical evidence, Chief Judge John J. Sirica of the U.S. District Court for the District of Columbia appointed a special Advisory Panel on White House Tapes to give expert advice to the court. The advisory panel consisted of six technical experts, jointly nominated by the counsel for the president and the special prosecutor, with the court's direction "… to study relevant aspects of the tape and the sounds recorded on it" [6]. The panel members included many familiar names from the engineering and acoustics communities: Richard H. Bolt (Bolt, Beranek, and Newman), Franklin S. Cooper (Haskins Laboratories), James L. Flanagan (AT&T Bell Laboratories), John G. McKnight (Magnetic Reference Laboratory), Thomas G. Stockham Jr. (University of Utah), and Mark R. Weiss (Federal Scientific Corp.).

The advisory panel performed a series of objective analyses of the tape itself, the magnetic signals on it, the electrical and acoustical signals generated by playback of the tape, and the properties of the recording equipment used to produce the magnetic signals on the tape. Analysis included observation of the audio signals as well as magnetic development of the domain patterns and head signatures on the tape. Ultimately the panel determined that the 18½-minute gap was due to several overlapping erasures performed with a specific model of tape recorder that differed from the device that produced the original recording. The panel's conclusion was based primarily on the characteristic start/stop magnetic signatures present on the subject tape.

The form of the panel's examination quickly became the standard approach for assessing the authenticity of forensic audio recordings:

1) Physically observe the entire length of the tape.
2) Document the total length and mechanical integrity of the tape, reels, and housing.
3) Verify that the recording is continuous with no unexplained stop/start sequences or erasures.
4) Perform critical listening of the entire tape.
5) Use nondestructive signal processing as needed for intelligibility enhancement.

Other well-known forensic audio cases include crime reconstruction attempts using acoustic evidence of a Dallas Police Department Dictaphone recording purportedly from the assassination of President Kennedy [7], interpretation of background sounds from cockpit voice recorder (black box) data [8], and the use of voice identification techniques for authenticating recordings of Osama bin Laden and other terrorists [9].

THERE IS POTENTIALLY MUCH TO BE GAINED BY APPLYING MODERN DSP THEORY TO PROBLEMS OF INTEREST TO THE FORENSICS COMMUNITY, BUT EXPERTS MUST UNDERSTAND THE STANDARDS AND PRACTICES OF THE LEGAL SYSTEM.

AUTHENTICITY

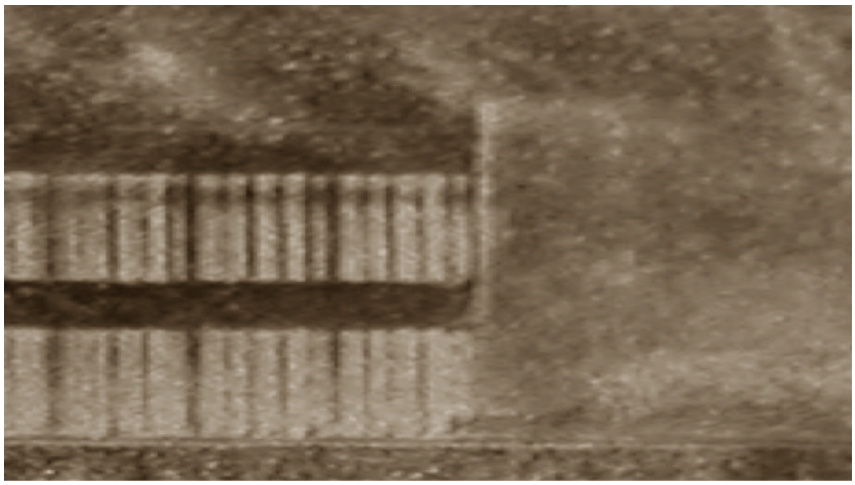
Like other types of forensic evidence, forensic audio may be subject to accidental or deliberate tampering. The court must be convinced of the authenticity and integrity of the audio evidence. What is authenticity? How can it be demonstrated? How might DSP help?

Ideally, an audio recording made for forensic purposes will be produced with authenticity verification in mind. For example, the recording will include an audio slate, consisting of a spoken announcement of the relevant information governing the recording, such as location, date, time, participants, model, and serial number of the recorder and the recording medium, etc. The recording should be made in one continuous session without pauses or stop/start sequences. Authenticity is also more verifiable if one deliberately allows uniquely identifiable background sounds such as clock tower chimes or radio broadcasts to be included in the audio recording.

FORENSIC AUDIO RECORDINGS

The authenticity of audio forensic evidence has traditionally focused on analog magnetic tape recordings. The forensic examiner performs a series of observations and tests to evaluate the integrity of the recording [2], [5], [10]. Despite the fact that digital recorders with solid-state flash memory are increasingly available, many law enforcement agencies in the United States still rely almost exclusively on analog cassette and microcassette recorders due to the agencies' inventory and familiarity with these devices—as well as authenticity concerns with digital data, as noted later in this section. The forensic audio authenticity

procedure for media such as analog tape typically follows the strategy employed by the Advisory Panel on White House Tapes. The methodology requires the examiner to observe the physical integrity of the recording medium, the quality of the recorded audio, and the consistency of the magnetic signatures present on the tape. The details of the process are described next.



[FIG1] **Magnetic development image of a record-over signature from a segment of analog magnetic tape. The two striped bars on the left side of the figure are the magnetic traces of a two-channel recording. The vertical line near the middle and the smudged image to the right are due to the magnetic erase head being energized and destroying the underlying magnetic patterns previously recorded. A continuous, unaltered segment of magnetic tape would not exhibit this erasure signature, so in this case the forensic examiner would suspect the tape had been deliberately altered after the original recording was made (from [11]).**

PHYSICAL HANDLING AND INSPECTION

The examiner documents the condition and properties of the evidentiary recording, including the length and condition of the tape, the condition of reels and housing, any manufacturing serial numbers or batch numbers, and the magnetic configuration on the tape (number of tracks, mono or stereo, etc.). The tape itself is inspected to look for any splices or other changes, and the recorder used to produce the tape is also inspected and tested.

CRITICAL LISTENING

The examiner carefully listens to the entire recording and notes any apparent alterations or irregularities. Any audible evidence of edits, splices, or audible discontinuities in background sounds, buzzes, tones, and so on are noted.

MAGNETIC SIGNATURE AND WAVEFORM OBSERVATIONS

The condition of the magnetic signals on the tape is examined using magnetic development techniques and compared to reference signatures of recordings obtained from the same recording device. An example magnetic development image is shown in Figure 1. The magnetic signatures associated with the record and erase heads, as well as the transitions from stop to record, record to pause, overrecording, and so forth, are examined for consistency. The examiner also observes and measures the electrical waveforms obtained during playback of the tape.

REPORT PREPARATION

Finally, the examiner analyzes the observations and writes a report explaining whether the tape is believed to be authentic, a copy, or altered in any manner after the original recording was made.

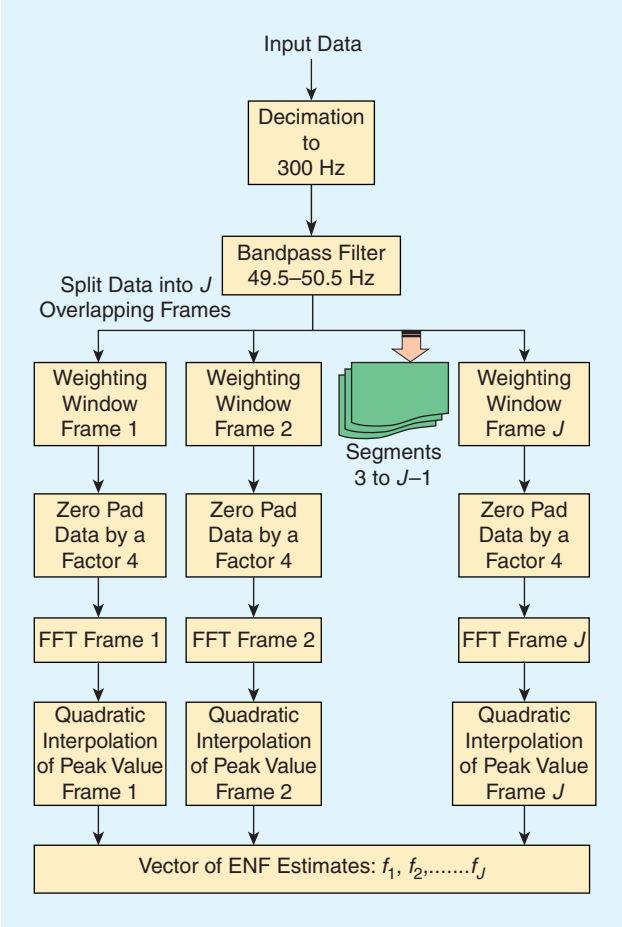
DIGITAL MEDIA

The question of authenticity becomes more complicated with digital recordings because the evidence of tampering or alterations is more difficult to discover than mechanical splices or overdubbing signatures in a physical master analog tape. A digital recording can be encoded with a checksum, processed with an embedded digital watermark, or otherwise encrypted, but it is difficult to exclude the possibility that the audio content itself was edited or manipulated prior to a surreptitious reencoding step. In general, ancillary information and meticulous chain-of-custody practices are essential.

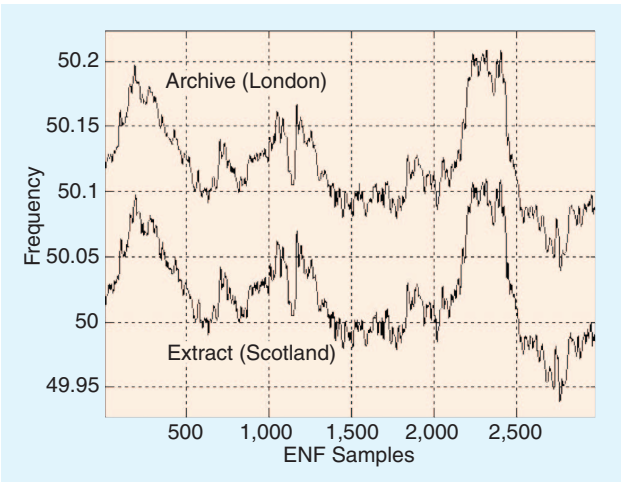
One recent effort in audio authentication that is applicable to digital recordings is analysis of residual signals due to coupling of the electrical power line frequency into the audio recording system [12–16]. The electric network frequency (ENF), nominally 60 Hz in the United States and 50 Hz in many other parts of the world, is not precisely constant but varies up to $+/-0.5$ Hz from time to time in an unpredictable fashion due to small mismatches between the electrical system load and system generation. The alternating magnetic field emanating from AC power lines can cause audible hum in the recording, which is usually considered undesirable. However, if hum is present, the characteristic frequency fluctuations can conceivably be traced to a particular date and time if sufficient data is available, since the ENF is consistent over the entire geographic region served by a synchronous AC network "grid." Thus, the ENF information extracted from a forensic recording could be compared to a database of known ENF data for the electrical power network to verify the date and time of the recording [15].

One system proposed for extracting the ENF from audio recordings is shown in Figure 2. A sample comparison of ENF data obtained from an audio recording and the reference ENF data from the electrical power system is shown in Figure 3.

The ENF procedure may not always be applicable in practice because magnetic field coupling from the power system may be minimal in well-designed audio equipment, when condenser or piezoelectric microphones are used, or when battery-operated equipment is used in an area away from the power grid [16].



[FIG2] Proposed ENF processing procedure (from [15]).



[FIG3] Automated match found using the techniques described. The extracted waveform has been offset by 0.1 Hz to aid visual comparison (from [15]).

Although no authoritative archive of ENF data is yet maintained throughout the world, various jurisdictions are looking at the feasibility of creating and storing their own records for possible future use [15].

ENHANCEMENT

Forensic audio recordings typically suffer from noise, distortion, interfering sounds, and other signal processing challenges that can impede proper analysis.

Perhaps the most common enhancement issue for forensic audio examiners is the clandestine surveillance recording of a relevant conversation via a hidden microphone. The surreptitious nature of the recording system often leads to poor microphone placement with respect to the participants, interference from wind and other environmental sounds, and rubbing or muffling due to clothing that comes in contact with the microphone.

When a recorded audio signal contains unwanted additive noise, it is desirable to enhance the perceived signal-to-noise ratio before playback (see [4], [17]–[25]). The enhancement process is generally performed iteratively off-line using a digital copy of the original evidentiary recording so that the original evidence is maintained unaltered.

The forensic enhancement procedure is nearly always performed on monophonic data as a blind (single-ended) process, since the only available data consist of the noise-degraded signal itself. The enhancement process must therefore be flexible and adaptive so that the examiner can deal with time-varying interference and acoustic corruption. Most examiners use processing in both the time domain and the frequency domain in an effort to supply the listener (stenographer, judge, or jury) with a signal that is of higher quality or intelligibility than the noisy original signal.

The goal of forensic audio enhancement may either be to improve intelligibility and reduce listener fatigue for speech transcription or to help reveal subtle or idiosyncratic background sounds that may be important investigative clues. The examiner first performs critical listening on the entire recording, making notes regarding the timing and quality of recognizable sounds and extraneous noises, as well as assessing the overall sonic quality of the material. If the examiner determines that enhancement is necessary, a variety of audio DSP tools are brought to bear.

COMMON DSP METHODS

The principal audio forensic enhancement procedures include time-domain level detectors and frequency-domain filters.

TIME-DOMAIN LEVEL DETECTION

Time-domain enhancement treats the amplitude envelope of the recorded audio signal. One example is gain compression, whereby the overall level (loudness) of the signal is adjusted to be relatively constant: quiet passages are amplified and loud passages are attenuated or left alone.

The traditional time-domain method for noise reduction, either analog or digital, uses a specified signal level, or threshold, that indicates the likely presence of the desired signal. The threshold is set (usually manually) high enough that when the desired signal is absent (for example, when there is a pause between sentences or messages), there is no background hiss or other noise. The threshold, however, must not be set so high

that the desired signal is affected when it is present. If the received signal is below the threshold, it is presumed to contain only noise, and the output signal level is reduced, or gated, accordingly. As used in this context, the term "gated" means that the signal is not allowed to pass through. By continuously monitoring the input signal level as compared with the threshold level, the time-domain level detection method gates the output signal on and off as the input signal level varies. In different contexts, this type of time-domain level detection system is referred to as squelch control, dynamic range expander, or noise gate. This process can make the received signal sound somewhat less noisy because the hiss goes away during the pause between words or sentences, but it is not particularly effective, in general, because it does nothing to reduce background noise when the gate is open and the desired signal is assumed to be present.

THE GOAL OF FORENSIC AUDIO ENHANCEMENT MAY EITHER BE TO IMPROVE INTELLIGIBILITY AND REDUCE LISTENER FATIGUE FOR SPEECH TRANSCRIPTION OR TO HELP REVEAL SUBTLE OR IDIOSYNCRATIC BACKGROUND SOUNDS THAT MAY BE IMPORTANT INVESTIGATIVE CLUES.

A discrete-time signal compressor/expander block diagram is shown in Figure 4. The block labeled "level detector" is typically an amplitude envelope detector or peak follower. One approach is to use a nonlinear full-wave comparison, such as

$$\begin{aligned} &\text{if } (|x[n]| > c[n-1]) \quad c[n] = \alpha\, c[n-1] \\ &\text{else } \; c[n] = \beta\, c[n-1], \end{aligned} \tag{1}$$

where α is the attack coefficient chosen to give the desired tracking when the input level is increasing and β is the decay coefficient chosen to follow the declining signal envelope. Thus, it is typical to choose $\alpha > 1$ and $0 < \beta < 1$.

The gain threshold, c_0, determines the level at which the gain control function becomes effective. For example, if gain expansion (squelch, or gain reduction at low levels) is needed, a gain calculation such as

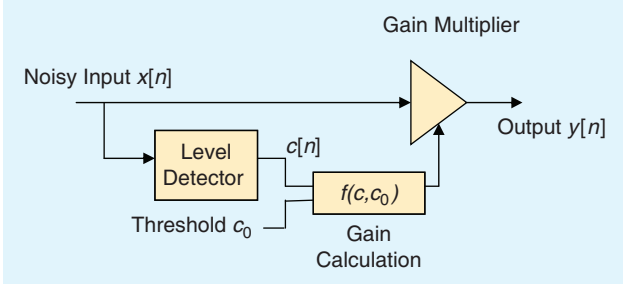
$$f(c) = \begin{cases} 1, & \text{if } c \geq c_0 \\ (c/c_0)^{\rho-1}, & \text{if } c < c_0 \end{cases} \tag{2}$$

could be used [26]. The parameter ρ defines the expansion factor: $\rho > 1$ causes $f(c)$ to be reduced when the detected level $c[n]$ is less than the threshold c_0. The larger the value of ρ, the more abrupt the gain change at low levels. A complementary approach is used to obtain gain compression or limiting (gain reduction at high levels).

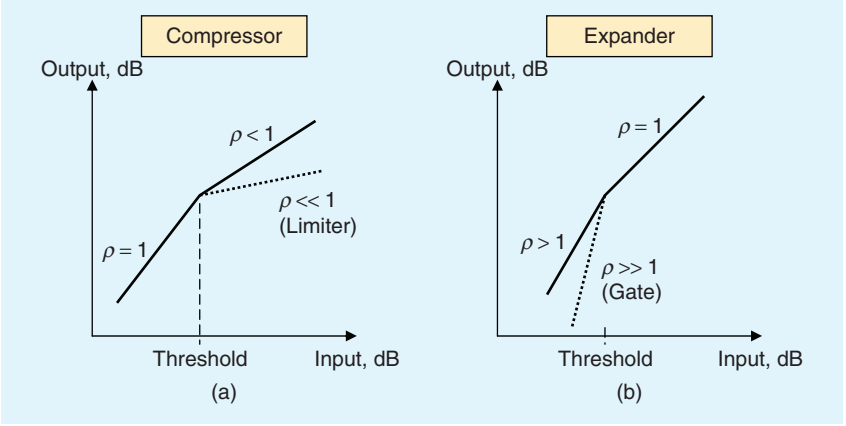
The usual way to depict gain compression/expansion is with a graph showing output versus input level, as shown in Figure 5. Note that the compression/expansion curves show the output level adjustment with respect to the input envelope level, not the instantaneous input sample value.

Level detection for segregating "noise only" segments from "signal plus noise" segments in forensic audio recordings is not particularly effective except when long periods of background noise are only intermittently interrupted by the desired signal, such as a recorded conversation in a noisy environment with lengthy pauses and gaps. Simple threshold gating does nothing to remove noise when the desired signal is present: the gate is simply "open" when the threshold is exceeded. Furthermore, the gate may open if there is a sudden burst of noise, a click, or some other loud sound that causes the signal level to exceed the threshold. In that case, the output signal quality is good only if the signal is sufficiently strong to mask the presence of the noise.

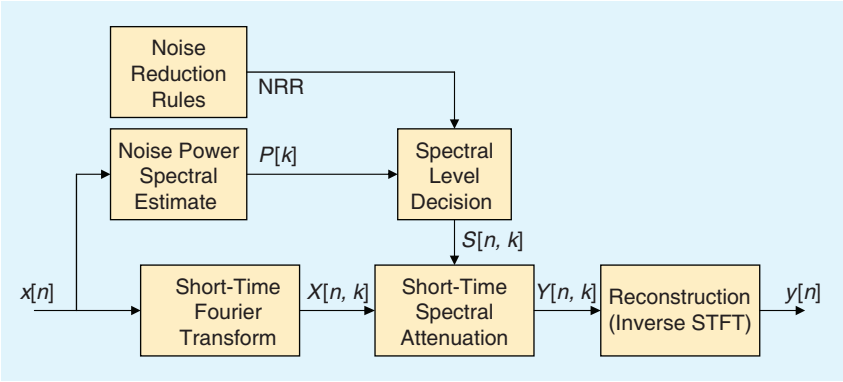
Another practical consideration is that changing the gain between the "gate open" mode and the "gate closed" mode must be done carefully to avoid audible noise modulation effects. The term gain pumping is used by recording engineers and refers to the audible sound of the noise appearing when the gate opens and then disappearing when the gate closes. Nevertheless, the time-domain gain compressor and expander functions can sometimes be useful in forensic enhancement situations.



[FIG4] Block diagram of a basic gain compressor/expander.



[FIG5] Depiction of gain (a) compression and (b) expansion behavior.

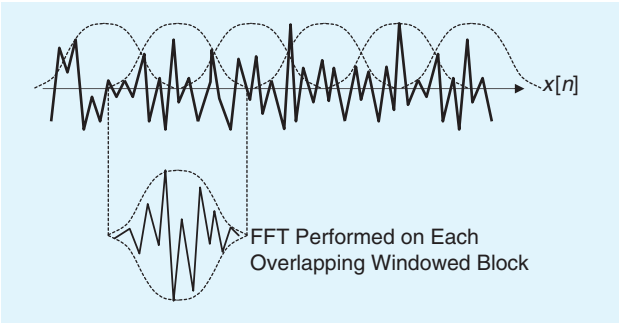


[FIG6] **Spectral noise reduction system framework (used with permission from [4]).**

The effectiveness of the time-domain methods can be improved using digital signal processing. Carefully controlling the attack and release times of the gate (i.e., how rapidly the processor responds to changes in the input signal) can minimize artifacts that would otherwise be confusing to the examiner. Other DSP improvements include look-ahead control software to cause the threshold to vary automatically if the noise level changes and splitting the gating decision into two or more frequency subbands. Using multiple frequency bands with individual gates means that the threshold can be set more optimally if the noise varies from one frequency band to another. For example, if the noise is mostly a low-frequency rumble or hum, the threshold can be set high enough to remove the noise in the low-frequency band while still maintaining a lower threshold in the high-frequency ranges. Despite these improvements, the time-domain processing methods for forensic audio enhancement are still limited because the processor cannot distinguish between noise and the desired signal other than on the basis of absolute signal envelope level.

FREQUENCY-DOMAIN FILTRATION

Frequency-domain methods for forensic audio enhancement often use some form of spectral subtraction. As its name implies, spectral subtraction involves forming an estimate of the noise spectrum (noise power as a function of frequency) and then subtracting this estimate from the noisy input signal spectrum.



[FIG7] **Short-time Fourier transform using overlapping block-based processing.**

The noise-reduced output is created by reconstructing the signal from the subtracted spectrum. Ideally, all the spectral energy below the noise estimate threshold is removed, so if the desired signal components exceed the noise level over much of the frequency range and if the noise estimate is sufficiently accurate, the technique can be useful and effective [20], [21].

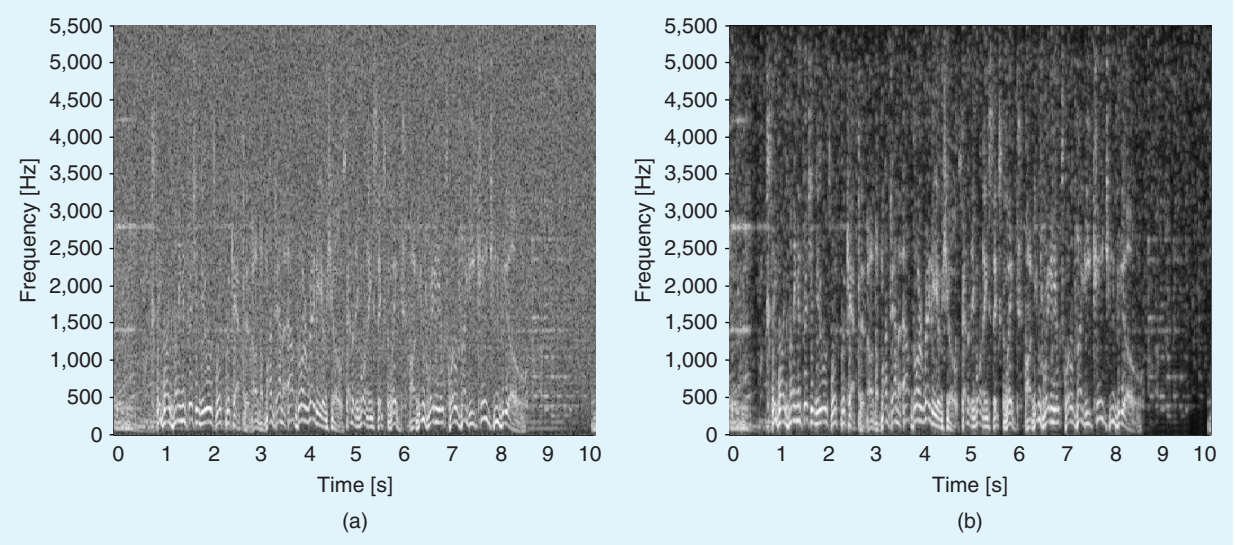
Unfortunately, if the actual noise level differs from the estimated noise spectrum, the noise reduction is incomplete and prone to undesirable audio artifacts. The residual spectral energy near the noise threshold can be audible as a whistling, tinkling sound that is sometimes referred to as birdie noise or musical noise [22–25]. Practical spectral subtraction systems allow frequent updates to the noise level estimate and incorporate techniques for reducing the residual musical noise. Thus, the quality and effectiveness of the spectral subtraction technique depends on the particular forensic task to be accomplished and the corresponding processing requirements.

The basic framework for spectral noise reduction is depicted in Figure 6. The short-time Fourier transform (STFT) produces a time sequence of spectral frames, or snapshots, $X[n, k]$, usually with a hop between the overlapping frames, as shown in Figure 7 [27], [28]. Within each frame, the spectral noise reduction algorithm makes decisions about which components are likely to be the desired signal and which are attributed to noise.

A hybrid single-ended noise reduction method extends the time-domain level detection and the frequency-domain spectral subtraction concepts by providing a means of distinguishing between the coherent behavior of the desired signal components and the incoherent (uncorrelated) behavior of the additive noise [4], [24], [25], [29]. The procedure identifies features that behave consistently over a short time window and attenuates or removes features that exhibit random or inconsistent fluctuations. As a single-ended method, the determination of noise versus signal cannot be perfect, but for many important forensic signals (such as noisy speech) the process can be made sufficiently reliable to improve the output signal for subsequent analysis [3].

One problem with simply applying spectral noise filtration is that common signals such as human speech contain noisy fricative and plosive components that are critical to speech intelligibility. Iterative pattern detection processing is often employed so that the fricative components are allowed primarily at boundaries between intervals with no voiced signal present and intervals with voiced components, since the presence and audibility of prefix and suffix consonant phonemes is a key feature for speech recognition [25].

An example of time-variant spectral (frequency-domain) noise reduction is shown in Figure 8.



[FIG8] Spectral noise reduction example. (a) STFT magnitude plot of noisy speech. The low contrast and gray shading reveal the broadband background noise. (b) STFT magnitude of enhanced version of the speech from (a). The higher contrast between the desired signal components (white) and the background level (black) show qualitatively the improved signal-to-noise ratio (from [25]).

INTERPRETATION

At the conclusion of a forensic audio assignment, the sponsor usually requires the examiner to prepare a report describing and interpreting the evidence, the methods employed, and the statistical basis for any opinions rendered.

It should be noted that although the performance of contemporary automatic speech transcription and speaker recognition systems is improving, there are no current court cases in the United States in which a judge has admitted computer-based transcription and recognition evidence. The challenge for audio enhancement and speech intelligibility research is to demonstrate performance that is appropriate for the standards necessary for a court to determine guilt or innocence in a criminal proceeding, and U.S. courts have always deferred to human expert interpretation, as described below. Here are several examples of the interpretation phase of audio forensics projects.

AUDIO FORENSICS REFERS TO THE ACQUISITION, ANALYSIS, AND EVALUATION OF AUDIO RECORDINGS THAT MAY ULTIMATELY BE PRESENTED AS ADMISSIBLE EVIDENCE IN A COURT OF LAW OR SOME OTHER OFFICIAL VENUE.

AURAL-SPECTROGRAPHIC VOICE IDENTIFICATION

Audio forensic examination of recorded dialog may lead to a legal dispute over the identity of one or more of the conversation participants. A criminal suspect or a party to civil litigation may deny being the individual who uttered the recorded words, especially if the recording was made via telephone without eyewitnesses to identify the talker visually. In these situations, the forensic audio examiner may be asked to identify or to exclude that the suspect was the source of the words in the recording in question.

The aural-spectrographic method for audio forensic voice identification is based on the judgment of a trained examiner who compares the unknown example of speech with one or more known examples [30]–[33]. As the name of the method implies, the task of the examiner is to render a judgment based on both an aural comparison (careful listening) and a visual comparison of speech spectrograms.

In a typical case, the examiner begins by listening critically to the recording of the unknown talker and identifies specific phrases that are distinctive and relatively noise-free. The examiner then arranges a recording session with the suspect to create exemplars that match the selected phrases of the unknown talker in pace, emphasis, and enunciation. The suspect repeats each example phrase multiple times to produce recordings with as close a match as possible to the timing and speech pattern of the unknown examples.

The examiner then generates individual audio files containing the unknown and exemplar utterances for each distinct phrase. The files are used for aural "A-B" comparisons and also to create spectrograms for visual comparison of formant shapes, discrete spectral features, and other patterns.

Although the aural-spectrographic examiner may use signal processing for enhancement and spectral display, the aural and visual observations ultimately lead to the examiner's opinion about the overall likelihood that the exemplars match or do not match the unknown recording. Specifically, the examiner reports one of the following decisions:

1) positive identification (the unknown speech positively matches the suspect exemplar)
2) probable identification
3) no decision
4) probable elimination
5) positive elimination (the suspect exemplar positively does not match the unknown speech).

The reliance of the aural-spectrographic examiner on his or her prior experience with subjective pattern matching has led to considerable controversy over the years regarding the veracity and scientific basis for the examination and therefore its admissibility in court [31]. At present, there are no generally accepted scientific studies that fully quantify the expected error rate of aural-spectrographic analysis, so courts must evaluate the admissibility of such expert testimony on a case-by-case basis [32], [33].

ONE PROBLEM WITH SIMPLY APPLYING SPECTRAL NOISE FILTRATION IS THAT COMMON SIGNALS SUCH AS HUMAN SPEECH CONTAIN NOISY FRICATIVE AND PLOSIVE COMPONENTS THAT ARE CRITICAL TO SPEECH INTELLIGIBILITY.

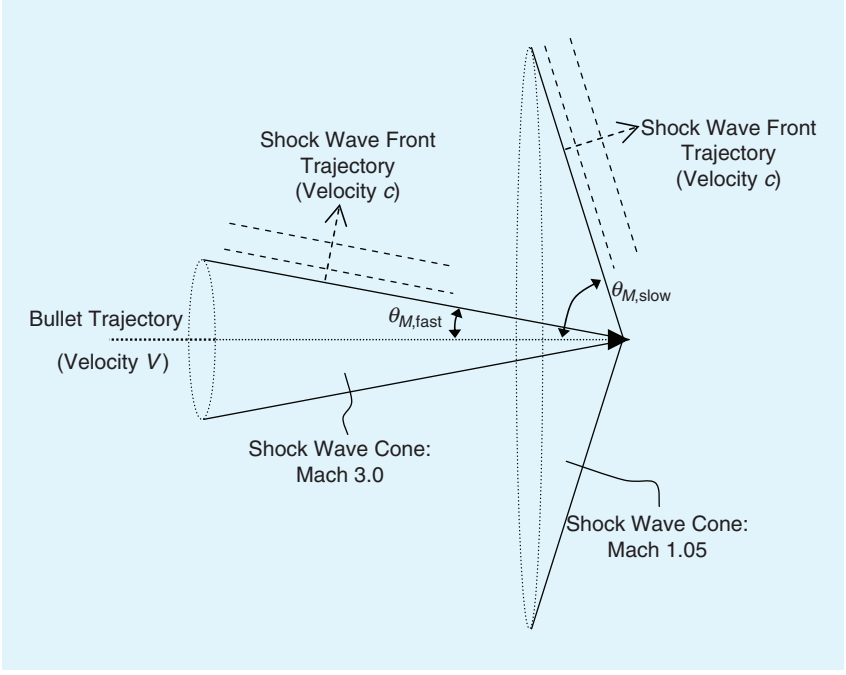
AIRCRAFT ACCIDENT INVESTIGATIONS

Modern commercial passenger aircraft and some military, corporate, and private planes are equipped with a flight data recorder (FDR) and a cockpit voice recorder (CVR). The FDR maintains a record of flight parameters such as time of day, altitude, aircraft orientation, airspeed, and so on. The CVR has audio channels to record radio communications and a cockpit area microphone located in the overhead panel above the pilot seats to pick up conversations and background sounds. The FDR maintains a record at least 25 h long, while the CVR typically records a loop 30–120 min in duration so that at least the last 30 min of cockpit sounds are documented in the event of a crash or other safety incident. The transcript of flight crew conversation plays an important role in the investigation of aircraft accidents [8].

In addition to the recorded dialog, the CVR inherently gathers other nonspeech information, including audible warning and alert signals, mechanical noises from the air frame, and the sound from the aircraft's engines. The CVR data can even help accident investigators identify crew respiration rates and other subtle clues about crew stress, exertion, and other situational information.

In one significant case involving audio forensic investigation using CVR data, examiners from the U.S. National Transportation Safety Board used a careful analysis of audio CVR material from the September 1994 crash near Pittsburgh of USAir Flight 427 (a Boeing 737 aircraft) to understand the behavior of the aircraft's engines and the timing, reactions, and efforts of the pilot and first officer during the incident. Among other details, the investigation included experiments to determine the ability of the cockpit microphone to pick up sound through structure-borne vibration [4].

A 1997 investigation of the CVR data from a Beechcraft 1900C commuter aircraft accident that occurred in 1992 used signal characteristics from both the cabin microphone and an unused CVR channel to study the theory that an in-flight engine separation was preceded by evidence of propeller whirl flutter attributable to a cracked truss in the engine mount [35].



[FIG9] Shock wave behavior for relatively fast (Mach 3.0) and slow (Mach 1.05) supersonic projectiles (used with permission from [38]).

GUNSHOT ACOUSTICAL ANALYSIS

Audio forensic analysis of recorded gunshots can help verify eyewitness (and earwitness) accounts and aid in crime scene reconstruction. The audio evidence can include the muzzle blast, the shock wave signature if the projectile is traveling at supersonic speed, the arrival of reflected and reverberated sound from nearby obstacles, and possibly even the characteristic sound of the firearm's mechanical action if the recording is obtained close to the shooting position [35]—[38].

The potential forensic audio evidence associated with gunshots can include several components, as described next.

MUZZLE BLAST

A conventional gun uses the rapid combustion of gunpowder to propel the bullet out of the firearm. The rapid ejection of

hot gas from the muzzle of the firearm causes an acoustic shock wave and chaotic noise referred to as the muzzle blast. The muzzle blast typically lasts for only a few milliseconds. If a recording microphone is located close to the gun barrel, the direct sound of the muzzle blast is the primary acoustical signal and typically overloads the microphone and/or the recording electronics due to the very high sound-pressure level. Microphones located at greater distances will typically exhibit the presence of multipath reflections, reverberation, and other effects of the sonic environment.

MECHANICAL ACTION

The mechanical action of the firearm includes the sound of the cocking and firing mechanism, the positioning of new ammunition by the gun's automatic or manual loading system, and possibly the telltale sounds of the spent cartridge being ejected and striking the ground. These mechanical sounds are very subtle compared to the high sound-pressure levels of the muzzle blast, so detection of the mechanical action may only be possible with closely miked recordings made near the shooter.

SINCE THE 1960s, THE U.S. FEDERAL BUREAU OF INVESTIGATION HAS CONDUCTED EXAMINATION OF AUDIO RECORDINGS FOR SPEECH INTELLIGIBILITY ENHANCEMENT AND AUTHENTICATION.

SUPERSONIC PROJECTILE

In addition to the muzzle blast and mechanical action, a third source of acoustic gunshot information is present if the bullet travels at supersonic speed [35], [38]. The projectile's speed, V, depends on the size of the charge, the mass of the bullet, and other ballistic factors. A supersonic bullet causes a characteristic shock wave pattern as it travels through the air. The shock wave expands as a cone behind the bullet, with the shock wave front propagating outward at the speed of sound, denoted by c. The shock wave cone has an inner angle θ_M that is related to the speed of the bullet by the formula $\theta_M = \arcsin(1/M)$, where $M = V/c$ is the Mach number. The shock wave geometry is shown in Figure 9.

A high-velocity bullet with V much greater than c results in a Mach number much greater than unity and thus a narrow shock wave cone angle, with the shock wave front propagating outward nearly perpendicularly to the bullet's path. For example, a bullet traveling at 3,000 ft/s at room temperature has $M = 2.67$, giving $\theta_M = \sim 22°$. On the other hand, if the bullet is traveling just barely over the speed of sound, M is approximately unity and the Mach angle is nearly 90°, meaning that the shock wave propagation is essentially parallel to the firing trajectory. Furthermore, the bullet will decelerate as it travels due to friction with the air, causing the Mach angle to broaden as the bullet slows downrange. Thus, forensic prediction of the shooting position and orientation based on the relative arrival timing of the shock wave and muzzle blast (and reflections from the surroundings) must take into account the estimated Mach number along the bullet's path.

GUNSHOT EXAMINATION ISSUES

The audio forensic challenges of gunshot analysis are tied to the impulsive nature of the sonic signatures for both the muzzle blast and the projectile's shock wave, if present. Although the gunshot sounds used in Hollywood movies and video game soundtracks are usually hundreds of milliseconds in duration, the actual duration of the muzzle blast is typically only 1–3 ms, while the shock wave over- and underpressure signature is just a few hundred microseconds in duration. Indeed, from a forensics standpoint the sound effects library gunshot recordings tell more about the acoustical impulse response of the surroundings than they do about the firearm itself, because such recordings deliberately contain an artificially high level of echoes and reverberation to enhance the emotional impact. In fact, earwitnesses who hear true gunfire often remark that the sounds seem like mere "pops" or "firecrackers" rather than gunshots, at least in comparison to their media-influenced expectations. Even if reverberation is not added deliberately, gunshot recordings obtained in acoustically reflective areas, such as indoors or outdoors in an urban area, may contain a mixture of overlapping shots and echoes that can complicate the analysis process.

The peak sound-pressure levels near the firearm can exceed 150 dB re 20 μPa. The high peak pressures associated with the gunshot sounds can cause clipping in the microphone and the input stage, and the extremely rapid rise times are usually sufficiently distorted by the recording system to make quantitative observation difficult. This is particularly true for recordings obtained via telephone.

CONCLUSIONS

This article has presented an overview of current practices in the field of audio forensics. Signal processing experts can and should look at forensic audio applications for their research and development efforts, but with the understanding that the peculiar requirements and demands of the criminal court system may lead to some frustration until new techniques are evaluated and accepted as admissible.

There is clearly a need for ongoing education of the courts and the public—who make up the jury pool—when considering the strengths and weaknesses of forensic audio material. There are increasingly frequent anecdotes in the forensic audio community regarding the so-called "*CSI* effect," referring to the fictional entertainment drama *Crime Scene Investigation* on U.S. broadcast television. Judges and jury members who are familiar with the *CSI* television series may come to court with expectations about the capabilities of signal enhancement and voice identification that are wholly unsupported in reality. Whether or not these misconceptions will lead to issues for jury deliberations remains to be seen.

IEEE Signal Processing Society members are encouraged to become informed about both the legal and the technical

challenges of audio forensics so that law enforcement, criminal justice, and accident investigation professionals are given the education and tools necessary to carry out their important work in contemporary society.

ACKNOWLEDGMENTS

The author gratefully acknowledges the comments of the anonymous reviewers and helpful conversations and guidance from Durand Begault, Christoph Musialik, Richard Sanders, and Steven R. Shaw.

AUTHOR

Robert C. Maher (rob.maher@montana.edu) received a B.S. degree from Washington University–St. Louis (1984), an M.S. degree from the University of Wisconsin–Madison (1985), and a Ph.D. from the University of Illinois at Urbana-Champaign (1989), all in the field of electrical engineering. From 1989 to 1996, he was a faculty member with the Department of Electrical Engineering, University of Nebraska–Lincoln. In 1996 he moved to Boulder, Colorado, to accept the position of Vice-President for Engineering with EuPhonics, Inc., (1996–1998), Engineering Manager for Audio Products with 3Com/U.S. Robotics (1998–2001), and adjunct professor with the University of Colorado–Boulder (2000–2002). In 2002, he joined the Department of Electrical and Computer Engineering, Montana State University–Bozeman, where he is currently department head and professor. His research, teaching, and consulting work involves the application of DSP to problems in audio engineering, acoustics, environmental sound, and audio forensics.

REFERENCES

[1] B. E. Koenig, "Authentication of forensic audio recordings," *J. Audio Eng. Soc.*, vol. 38, no. 1/2, pp. 3–33, Jan./Feb. 1990.

[2] *AES Standard for Forensic Purposes—Criteria for the Authentication of Analog Audio Tape Recordings*, AES Standard 43-2000.

[3] B. E. Koenig, D. S. Lacey, and S. A. Killion, "Forensic enhancement of digital audio recordings," *J. Audio Eng. Soc.*, vol. 55, no. 5, pp. 252–371, May 2007.

[4] C. Musialik and U. Hatje, "Frequency-domain processors for efficient removal of noise and unwanted audio events," in *Proc. Audio Engineering Society 26th Conf., Audio Forensics in the Digital Age*, Denver, CO, July 2005, pp. 65–77.

[5] *AES Recommended Practice for Forensic Purposes—Managing Recorded Audio Materials Intended for Examination*, AES Standard 27-1996.

[6] Advisory Panel on White House Tapes, *"The Executive Office Building Tape of June 20, 1972: Report on a technical investigation*," United States District Court for the District of Columbia, May 31, 1974. [Online]. Available: http://www.aes.org/aeshc/docs/forensic.audio/watergate.tapes.report.pdf

[7] National Academy of Sciences, "*Report of the Committee on Ballistic Acoustics*," Washington, D.C.:National Academy Press, 1982. [Online]. Available: http://www.nap.edu/catalog.php?record_id=10264

[8] G. Byrne, *Flight 427: Anatomy of an Air Disaster*. New York: Springer-Verlag, 2002.

[9] J. S. Sachs. (2003, Mar.). Graphing the voice of terror. Popular Sci., pp. 38–43 [Online]. Available: http://www.popsci.com/scitech/article/2003-02/graphing-voice-terror

[10] Scientific Working Group on Digital Evidence. (2008, Jan. 31). *SWGDE Best Practices for Forensic Audio*, Version 1.0 [Online]. Available: http://www.swgde.org/documents/swgde2008/SWGDEBestPracticesforForensicAudioV1.0.pdf

[11] D.R. Begault, B.M. Brustad, and A.M. Stanley, "Tape analysis and authentication using multi-track recorders," in *Proc. Audio Eng. Soc. 26th Conf., Audio Forensics in the Digital Age*, Denver, CO, July 2005, pp. 115–121.

[12] C. Grigoras, "Digital audio recording analysis: The electric network frequency (ENF) criterion," *Int. J. Speech Language Law*, vol. 12, no. 1, pp. 63–76, 2005.

[13] E. B. Brixen, "Techniques for the authentication of digital audio recordings," in *Proc. Audio Engineering Society 122nd Conv.*, Vienna, Austria, 2007, Convention Paper 7014.

[14] C. Grigoras, "Application of ENF analysis method in authentication of digital audio and video recordings," in *Proc. Audio Engineering Society 123rd Conv.*, New York, 2007, Convention Paper 1273.

[15] A. J. Cooper, "The electric network frequency (ENF) as an aid to authenticating forensic digital audio recordings—An automated approach," in *Proc. Audio Engineering Society 33rd Conf., Audio Forensics—Theory and Practice*, Denver, CO, June 2008, pp. 1–10.

[16] E. B. Brixen, "ENF—Quantification of the magnetic field," in *Proc. Audio Engineering Society 33rd Conf., Audio Forensics—Theory and Practice,* Denver, CO, June 2008, pp. 1–6.

[17] M. R. Weiss, E. Aschkenasy, and T. W. Parsons, "*Study and development of the INTEL technique for improving speech intelligibility*," Nicolet Scientific Corp., Final Rep. NSC-FR/4023, 1974.

[18] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.

[19] M. R. Weiss and E. Aschkenasy, "*Wideband speech enhancement (addition)*," Final Tech. Rep. RADC-TR-81-53, DTIC ADA100462, 1981.

[20] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, no. 2, pp. 113–120, 1979.

[21] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, no. 2, pp. 137–145, 1980.

[22] O. Cappé, "Elimination of the musical noise phenomenon with the ephraim and malah noise suppressor," *IEEE Trans. Speech Audio Processing*, vol. 2, no. 2, pp. 345–349, 1994.

[23] D. E. Tsoukalas, J. N. Mourjopoulos, and G. Kokkinakis, "Speech enhancement based on audible noise suppression," *IEEE Trans. Speech Audio Processing*, vol. 5, no. 6, pp. 479–514, 1997.

[24] S. Godsill, P. Rayner, and O. Cappé, "Digital audio restoration," in *Applications of Digital Signal Processing to Audio and Acoustics*, M. Kahrs and K. Brandenburg, Eds. Norwell, MA: Kluwer, 1998, pp. 133–194.

[25] R.C. Maher, "Audio enhancement using nonlinear time-frequency filtering," in *Proc. Audio Engineering Society 26th Conf., Audio Forensics in the Digital Age*, Denver, CO, July 2005, pp. 104–112.

[26] S. J. Orfanidis, *Introduction to Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1996.

[27] J.B. Allen and L.R. Rabiner, "A unified approach to short time Fourier analysis and synthesis," *Proc. IEEE*, vol. 65, no. 11, pp. 1558–1564, 1977.

[28] T.F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*. Englewood Cliffs, NJ: Prentice-Hall, 2002.

[29] J. Moorer and M. Berger, "Linear-phase bandsplitting: Theory and applications," *J. Audio Eng. Soc.*, vol. 34, no. 3, pp. 143–152, 1986.

[30] R. H. Bolt, F. S. Cooper, E. E. David, P. B. Denes, J. M. Pickett, and K. N. Stevens, "Identification of a speaker by speech spectrograms," *Science*, vol. 166, no. 3903, pp. 338–342, 1969.

[31] R. H. Bolt, F. S. Cooper, E. E. David, P. B. Denes, J. M. Pickett, and K. N. Stevens, "Speaker identification by speech spectrograms: A scientist's view of its reliability for legal purposes," *J. Acoust. Soc. Amer.*, vol. 47, no. 2, pp. 597–612, 1970.

[32] National Academy of Sciences, Committee on Evaluation of Sound Spectrograms, "On the theory and practice of voice identification," Washington, D.C.: National Academy Press, Rep. 0-309-02973-16, 1979.

[33] F. Poza and D. R. Begault, "Voice identification and elimination using aural-spectrographic protocols," in *Proc. Audio Engineering Society 26th Conf., Audio Forensics in the Digital Age*, Denver, CO, July 2005, pp. 21–28.

[34] R. O. Stearman, G. H. Schulze, and S. M. Rohre, "Aircraft damage detection from acoustic and noise impressed signals found by a cockpit voice recorder," in *Proc. Nat. Conf. Noise Control Engineering*, vol. 1, 1997, pp. 513–518.

[35] B. M. Brustad and J. C. Freytag, "A survey of audio forensic gunshot investigations," in *Proc. Audio Engineering Society 26th Conf., Audio Forensics in the Digital Age*, Denver, CO, July 2005, pp. 131–134.

[36] R. C. Maher, "Modeling and signal processing of acoustic gunshot recordings," in *Proc. IEEE Signal Processing Society 12th DSP Workshop*, Jackson Lake, WY, Sept. 2006, pp. 257–261.

[37] R. C. Maher, "Acoustical characterization of gunshots," in *Proc. IEEE SAFE 2007: Workshop on Signal Processing Applications for Public Security and Forensics*, Washington, D.C., Apr. 2007, pp. 109–113.

[38] R. C. Maher and S. R. Shaw, "Deciphering gunshot recordings," in *Proc. Audio Engineering Society 33rd Conf., Audio Forensics—Theory and Practice*, Denver, CO, June 2008, pp. 1–8. [SP]